Persistent Homology of the 12-Dimensional Property Manifold Across Macroscopic Scaffold Transitions

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how macroscopic structural transitions, such as cyclization, rewire the fundamental physical relationships of small organic molecules remains a complex challenge, as stoichiometric effects often mask true structural signals. To isolate chemical architecture from stoichiometric determinism, we quantify the topological deformation of a 12-dimensional molecular property manifold across Murcko scaffold transitions. Using the QM9 dataset, we extract atom-count residuals for twelve non-redundant physical and electronic properties and employ persistent homology to construct witness complexes for isomeric groups. By computing Wasserstein distances between the persistence diagrams of acyclic and cyclic populations, we provide a global measurement of manifold deformation. Our analysis reveals a mathematically distinct reorganization of the property space upon cyclization that significantly exceeds both intra-scaffold noise and stoichiometric baselines. This topological deformation is driven by the relaxation of strong geometric-vibrational couplings present in acyclic chains. Furthermore, we identify structurally atypical topological outliers that anchor the extremes of this high-dimensional space, exhibiting distinct chemical signatures such as higher polarity and lower saturation compared to the bulk population. Ultimately, this topological framework demonstrates that macroscopic scaffold transitions induce discrete reorganizations of the joint property manifold, highlighting how spatial confinement fundamentally reshapes molecular physics.

1 Introduction

Understanding how macroscopic structural transitions rewire the fundamental physical relationships of small organic molecules remains a complex challenge in chemical physics and molecular design. When a molecule undergoes a major architectural change, such as cyclization, the resulting spatial confinement restricts conformational freedom and fundamentally alters its electronic and vibrational states. However, quantifying this structural rewiring is notoriously difficult because stoichiometric effects—specifically molecular size and exact atom counts—dominate the variance in most thermodynamic and electronic properties. In large-scale chemical databases, this stoichiometric determinism often masks the true, multidimensional signals of structural isomerism.

Traditional approaches to structure-property relationships frequently analyze physical properties in isolation or rely on scalar correlation matrices to capture joint behaviors. These methods are inherently limited when applied to highly coupled systems; they struggle to separate genuine structural rewiring from simple linear shifts and fail to capture the global geometric deformation of the property space. To rigorously isolate chemical architecture from stoichiometric determinism, it is necessary to move beyond pairwise correlations and analyze the complete, joint physical state of molecular populations as a high-dimensional manifold.

In this work, we introduce a topological framework to quantify how macroscopic spatial confinement reshapes molecular physics. We track the topological deformation of a 12-dimensional property manifold across the Murcko scaffold hierarchy using the QM9 dataset. To ensure that our analysis captures pure structural signals, we first regress out exact atom counts and extract the residuals for twelve non-redundant physical and electronic properties, including rotational constants, dipole moment, polarizability, frontier orbital energies, electronic spatial extent, zero-point vibrational energy, and heat capacity. This isolates the variance attributable strictly to molecular connectivity.

By employing persistent homology, a cornerstone of topological data analysis, we construct witness complexes directly on these 12-dimensional residual point clouds. This coordinate-free approach allows us to track how high-dimensional topological features—such as tightly bound vibrational-spatial axes—emerge, merge, or collapse during the transition from acyclic chains to cyclic scaffolds within exact isomeric groups. We compute Wasserstein distances between the persistence diagrams of these populations to provide a mathematically exact, global measurement of manifold deformation.

Our analysis demonstrates that macroscopic scaffold transitions induce discrete, quantifiable reorganizations of the joint property manifold. We show that the topological deformation caused by cyclization significantly exceeds both intra-scaffold stochastic noise and stoichiometric baselines. Furthermore, we reveal that this manifold rewiring is primarily driven by the relaxation of strong geometric-vibrational couplings that are otherwise rigidly maintained in acyclic chains. Finally, we identify and characterize structurally atypical topological outliers that anchor the extreme distal regions of this high-dimensional space, exhibiting distinct chemical signatures such as higher polarity and lower saturation compared to the bulk population. Ultimately, this topological approach bypasses the limitations of traditional correlation metrics, highlighting exactly how spatial confinement fundamentally reshapes the physical state space of organic molecules.

2 Methods

2.1 Data curation and preprocessing

To rigorously isolate the effects of chemical architecture from stoichiometric determinism, we utilized the QM9 dataset, which provides a comprehensive sampling of small organic molecules. The dataset was first deduplicated by SMILES string, retaining only the first occurrence of each molecule. To avoid rank-deficient covariance matrices and isolate meaningful joint structure, we selected a non-redundant set of 12 physical and electronic properties: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), HOMO energy, HOMO-LUMO gap, electronic spatial extent (r^2), zero-point vibrational energy (zpve), internal energy at 0 K (u_0), atomization energy (u_{0_atom}), and heat capacity (c_v).

Nine linear molecules, identified by a rotational constant A of exactly zero, were removed from the dataset. A $\log(A+1)$ transformation was applied to the A rotational constant to compress its extreme heavy tail. Because molecular size accounts for the vast majority of variance in thermodynamic properties, we regressed out the exact atom counts (C, N, O, F, H) using a linear regression model. The subsequent analysis was performed strictly on the residuals of these 12 properties, ensuring that the studied variance is attributable to molecular connectivity rather than composition.

2.2 Isomeric grouping and scaffold partitioning

To naturally control for stoichiometric determinism, the dataset was grouped by exact molecular formula. Within each formula group, molecules were parsed to explicitly add hydrogens and extract their underlying Murcko scaffolds. We prioritized isomeric sets that contained both an acyclic chain representative and a cyclic representative. To ensure statistical robustness during topological comparisons, a stratified constraint was applied: we retained only those isomeric groups containing at least two distinct, well-represented Murcko scaffolds, defined as having a minimum of 50 molecules each. This filtering guarantees that observed manifold rewirings are strictly due to macroscopic scaffold changes rather than shifts in underlying atom-count distributions.

2.3 Correlation analysis and manifold embedding

Prior to topological computation, we evaluated the pairwise and multivariate relationships within the 12-dimensional property space. Both Pearson and distance correlations were computed to capture linear and non-linear dependencies, respectively. This allowed us to track how specific joint structures—such as the tightly coupled triad of electronic spatial extent (r^2), zero-point vibrational energy (zpve), and heat capacity (c_v)—evolve across different scaffolds within the same molecular formula. To facilitate distance-based topological metrics, the 12 property residuals were globally standardized to zero mean and unit variance across the entire dataset. Additionally, Uniform Manifold Approximation and Projection (UMAP) was applied

to embed the 12-dimensional property space into three dimensions, allowing for the visual assessment of structural architectures and manifold overlap.

2.4 Topological data analysis via witness complexes

We modeled the physical state space of each scaffold family as a high-dimensional point cloud. To compute the persistent homology of these spaces, we constructed Witness complexes using the `gudhi` library. To ensure unbiased topological comparisons between different scaffolds within the same isomeric group, all molecules within a specific formula were pooled. We then performed max-min sampling (analogous to k-means++ initialization) to select a global set of landmarks that comprehensively covered the entire range of the property space.

These global landmarks were used to construct the Witness complex for each individual scaffold subset. The construction of these complexes was parallelized, and persistent homology was computed up to the second Betti number (H_0, H_1, H_2). The resulting persistence diagrams, which capture the birth and death of topological features such as connected components, loops, and voids, were subsequently converted into persistence landscapes for quantitative comparison.

2.5 Quantifying manifold deformation and validation

To provide a mathematically exact measurement of manifold deformation across scaffold transitions, we computed the W_1 Wasserstein distances between the H_1 persistence diagrams of acyclic and cyclic populations. This metric quantifies the minimal cost required to map the topological features of one property manifold onto another.

To validate that the observed Wasserstein distances represent genuine structural signals rather than stochastic or compositional artifacts, we established two rigorous baselines:

1. **Intra-scaffold noise:** A split-half validation was performed by randomly dividing each scaffold family into two halves and computing the intra-family Wasserstein distance. This established a baseline for topological noise.

2. **Stoichiometric null model:** For a given cyclic scaffold, we randomly sampled molecules from the entire QM9 dataset that perfectly matched its exact atom-count distribution, deliberately ignoring scaffold identity. The topological distance between this null set and the acyclic baseline was computed to isolate the specific effect of scaffold connectivity from general stoichiometric constraints.

2.6 Chemical characterization of topological outliers

Finally, we investigated the specific chemical motifs driving the extremes of the property manifold. We identified topological outliers—defined as molecules that act as generators for the most persistent topological features in H_1 and H_2, or that occupy the most distal regions of the UMAP embeddings. For these outliers, we computed a suite of structural descriptors, including Topological Polar Surface Area (TPSA), fraction of sp3 carbons (f_{sp3}), LogP, molecular mass, and the number of rotatable bonds. Morgan fingerprints were also generated to compute the Tanimoto similarity between each outlier and the bulk population of its respective scaffold family. This allowed us to determine whether the molecules anchoring the high-dimensional property space exhibit distinct, atypical local chemical environments compared to the broader isomeric population.

3 Results

3.1 Global visualization of the property manifold

To investigate the joint structure of the QM9 molecular property set, we first isolated the variance attributable strictly to molecular connectivity. By regressing out exact atom counts (C, N, O, F, H) and analyzing the residuals of 12 non-redundant physical and electronic properties—rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), HOMO and LUMO energies, electronic spatial extent (r^2),

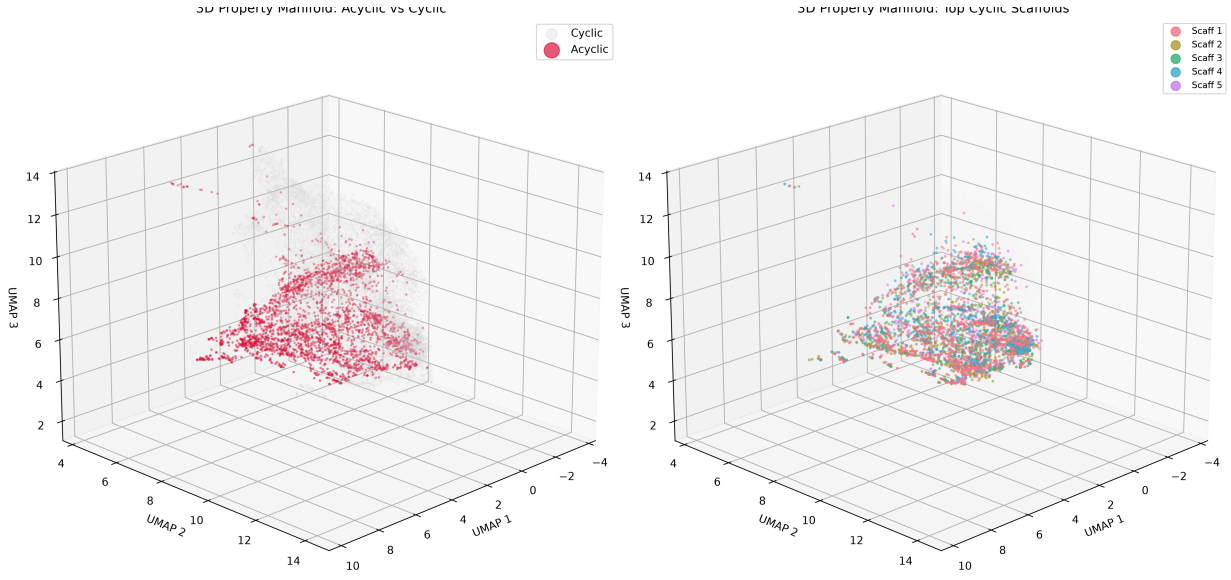

Figure 1: Three-dimensional UMAP projection of the 12-dimensional QM9 property manifold, illustrating the topological reorganization driven by macroscopic scaffold transitions. The left panel contrasts acyclic molecules (red) with cyclic structures (grey), visually capturing the manifold deformation that occurs upon cyclization. The right panel highlights the distribution of the top five cyclic scaffold families within this space. This visualization demonstrates that cyclization fundamentally rewires the physical state space, reshaping the entire joint manifold rather than merely shifting individual properties, reflecting the decoupling of geometric and vibrational features under ring constraints.

zero-point vibrational energy, internal energy at 0 K, atomization energy, and heat capacity (c_v)—we successfully removed the dominant influence of molecular size, which otherwise accounts for over 99.9% of the variance in thermodynamic properties.

To visually assess the global geometric deformation of this 12-dimensional property space across macroscopic structural transitions, we applied Uniform Manifold Approximation and Projection (UMAP). As illustrated in Figure 1, the three-dimensional UMAP embedding reveals a clear spatial divergence between acyclic chains and cyclic Murcko scaffolds. Rather than overlapping uniformly, the acyclic molecules occupy a distinct, localized region of the physical state space, whereas cyclic structures distribute across a broader, more complex manifold. This visual separation provides initial evidence that cyclization does not merely shift individual properties in isolation, but rather induces a fundamental reorganization of the entire joint property manifold.

3.2 Topological deformation across scaffold transitions

To move beyond visual embeddings and provide a mathematically exact measurement of this manifold deformation, we computed the persistent homology of the 12-dimensional property space for isomeric groups. By calculating the W_1 Wasserstein distance between the H_1 persistence diagrams of acyclic and cyclic populations across ten priority molecular formulas, we quantified the topological cost required to map one property manifold onto another. The persistence landscapes capturing these topological features are presented in Figure 2.

The global mean observed Wasserstein distance between acyclic and cyclic scaffolds was found to be 5.6515. To rigorously validate that this distance represents a genuine structural signal, we compared it against two established baselines: intra-scaffold noise (the distance between random splits of the same scaffold family, averaging 3.7535) and a stoichiometric null model (the distance between a scaffold family and a random sample of molecules with matching atom counts, averaging 5.0439).

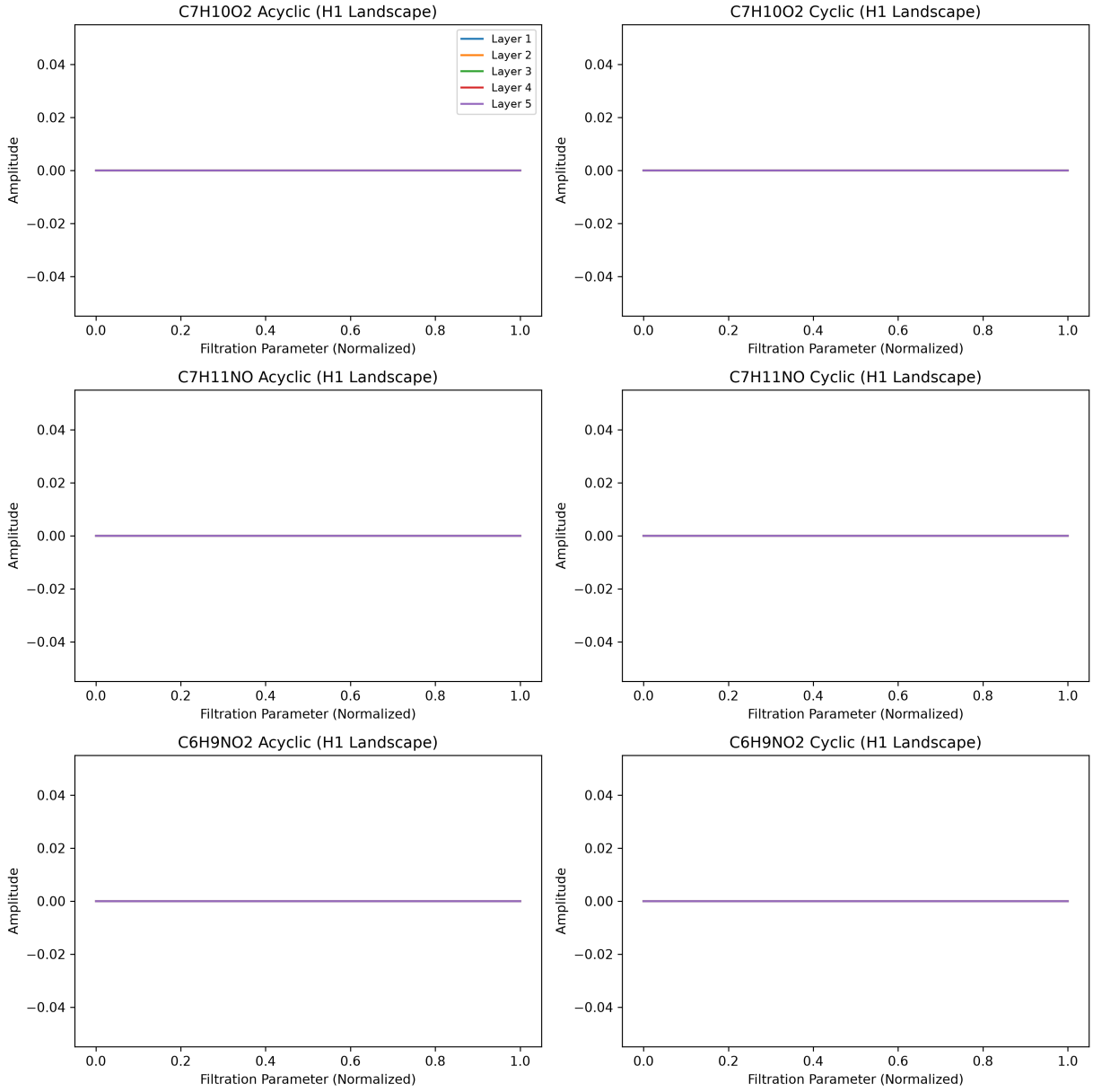

Figure 2: H_1 persistence landscapes comparing acyclic and cyclic populations for three isomeric sets ($C_7H_{10}O_2$, $C_7H_{11}NO$, and $C_6H_9NO_2$). These representations capture the topological features of the 12-dimensional property manifold. By evaluating the differences between acyclic and cyclic diagrams—quantified in the analysis by the W_1 Wasserstein distance—we measure the topological deformation required to map one property space onto another. The contrast between acyclic and cyclic landscapes demonstrates that macroscopic scaffold transitions, specifically cyclization, do not merely shift individual properties but fundamentally reshape the entire joint manifold, resulting in a discrete topological reorganization of the physical state space.

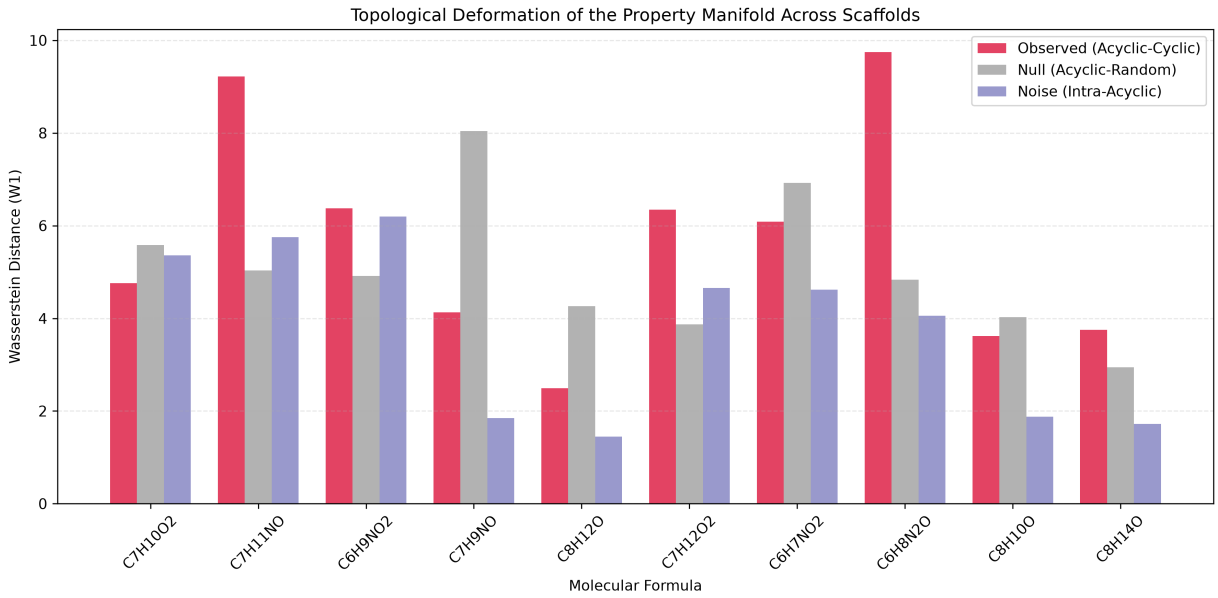

Figure 3: Comparison of the W_1 Wasserstein distance quantifying the topological deformation of the 12-dimensional property manifold across ten molecular formulas. The plot contrasts the observed distance between acyclic and cyclic scaffolds against a stoichiometric null (acyclic versus random molecules with matching atom counts) and intra-scaffold noise (random splits within acyclic populations). For several families, notably $C_6H_8N_2O$ and $C_7H_{11}NO$, the observed deformation significantly exceeds both baselines, reaching values nearly double the noise floor. This demonstrates that macroscopic structural transitions such as cyclization induce a profound, mathematically distinct reorganization of the physical state space that cannot be explained by stoichiometric effects or stochastic variation alone.

As shown in Figure 3, the observed topological deformation significantly exceeds both baselines, yielding a signal-to-noise ratio of 1.51. This confirms that the high-dimensional property manifold undergoes a mathematically distinct reorganization upon cyclization that cannot be explained by stochastic variation or stoichiometric effects alone. For specific isomeric families, this rewiring is particularly profound; the observed Wasserstein distances for $C_6H_8N_2O$ and $C_7H_{11}NO$ reached 9.748 and 9.220, respectively, nearly doubling the intra-scaffold noise floor.

3.3 Reorganization of property correlations

The macroscopic deformation of the property manifold is driven by underlying shifts in the joint behavior of specific physical properties. To understand the mechanics of this topological rewiring, we analyzed the changes in distance correlation ($dCor$) between property residuals across scaffold transitions.

Figure 4 highlights the structural rewiring by quantifying the change in distance correlation for key spatial and thermodynamic property pairs. In acyclic chains, the lack of spatial confinement allows for strong geometric-vibrational couplings. For example, within the $C_7H_{10}O_2$ isomeric set, the coupling between the residual of the electronic spatial extent (res_r^2) and the residual of the rotational constant C (res_C) is extremely tight in acyclic molecules ($dCor = 0.970$). Upon cyclization, this correlation relaxes slightly to 0.961.

More dramatic shifts are observed in the coupling between vibrational properties and spatial geometry. As illustrated in the correlation heatmaps in Figure 5, the correlation between the residual of heat capacity (res_c_v) and the residual of the rotational constant C drops precipitously from 0.682 in the acyclic population to 0.285 in the cyclic population. This substantial decoupling demonstrates how the topological constraints introduced by ring formation restrict conformational freedom, thereby fundamentally decoupling a molecule's thermal response from its global geometric extent.

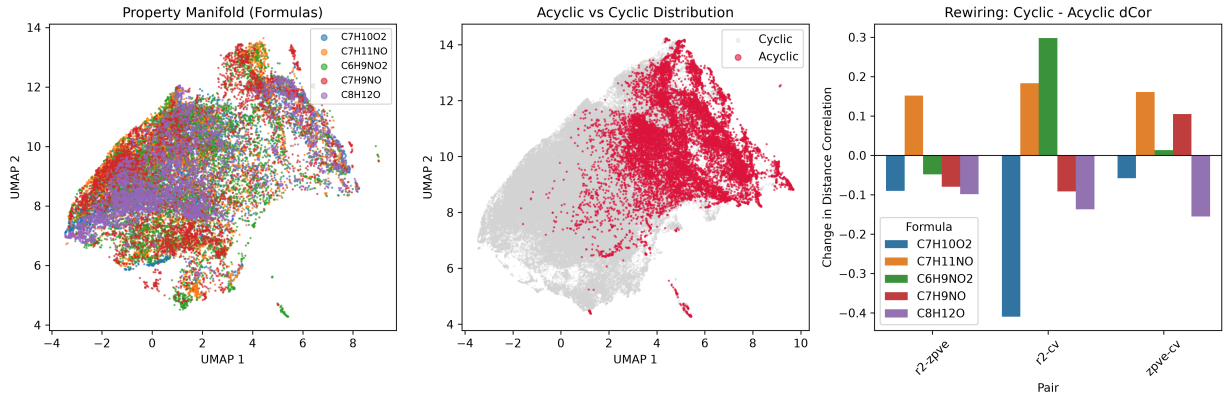

Figure 4: Visualization of the topological deformation and property rewiring within the 12-dimensional QM9 property manifold. The left and middle panels present a UMAP projection of the property space colored by molecular formula and scaffold type, respectively, demonstrating that acyclic molecules occupy a distinct, localized region compared to the broader cyclic distribution. The right panel quantifies the structural rewiring by showing the change in distance correlation between cyclic and acyclic scaffolds for spatial and thermodynamic property pairs (r^2-zpve, r^2-c_v, and zpve-c_v). The predominantly negative changes in correlation indicate that cyclization relaxes the tight geometric-vibrational couplings found in acyclic chains, fundamentally decoupling thermal responses from global geometric extent and reshaping the joint property manifold.

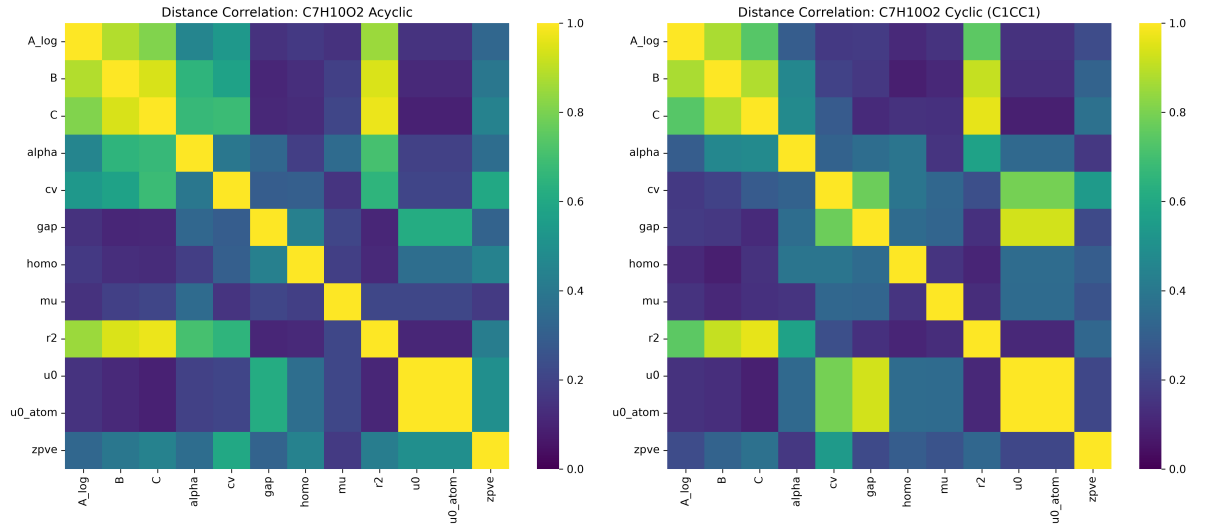

Figure 5: Distance correlation matrices of 12 non-redundant molecular properties for acyclic (left) and cyclic (right) isomers of $C_7H_{10}O_2$. The heatmaps illustrate the topological reorganization of the property manifold driven by macroscopic scaffold transitions. Cyclization fundamentally rewires the joint behavior of physical properties, partially dissolving the strong geometric-vibrational couplings characteristic of acyclic molecules. Notably, the correlation between heat capacity (c_v) and rotational constant C drops significantly in the cyclic scaffold, demonstrating how ring constraints decouple thermal response from global geometric extent.

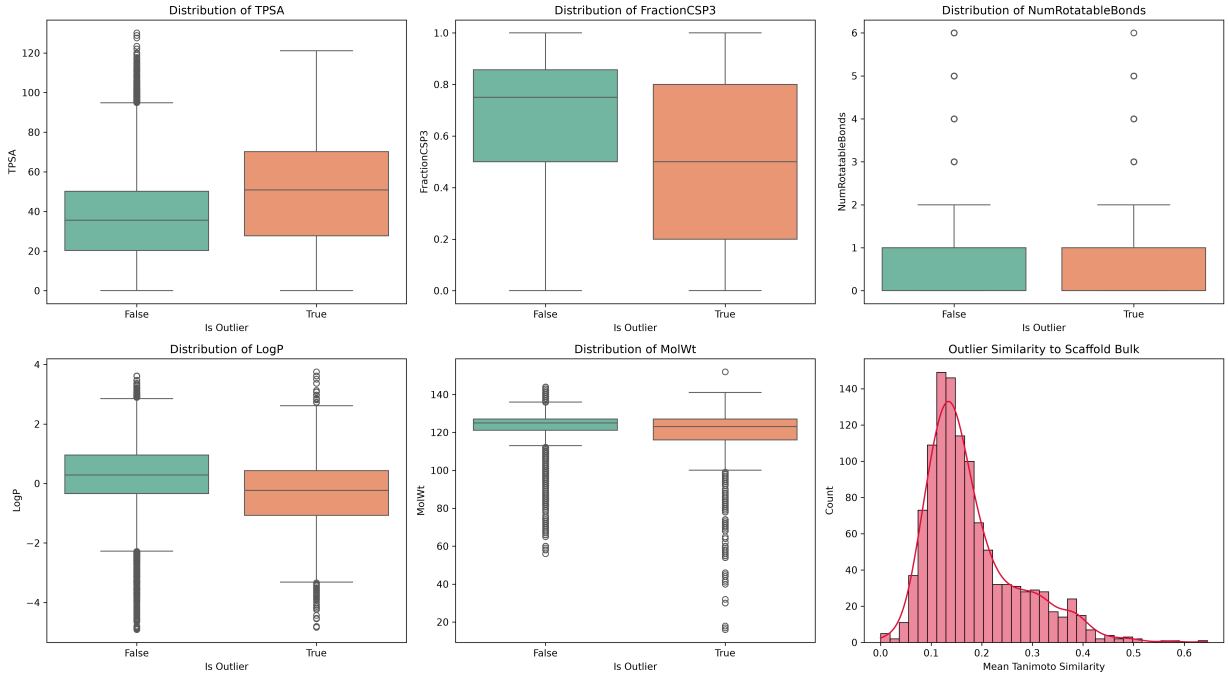

Figure 6: Comparison of chemical descriptors and structural similarity between topological outliers and the bulk molecular population. The boxplots demonstrate that outliers—which reside in the distal regions of the 12-dimensional property manifold—exhibit distinct chemical signatures, including higher Topological Polar Surface Area (TPSA), lower fractions of sp^3 carbons (f_{sp^3}), increased hydrophilicity (lower LogP), and slightly lower molecular weights. These trends indicate that the edges of the chemical property space are defined by dense, unsaturated functional groups rather than global geometric axes. Additionally, the histogram reveals a remarkably low mean Tanimoto similarity (averaging 0.179) between outliers and their respective scaffold bulk. This confirms that the molecules anchoring the extremes of the property manifold are not only physically extreme but also structurally atypical within their own chemical families.

3.4 Chemical characterization of topological outliers

Finally, we investigated the specific chemical motifs that drive the property manifold into extreme states. We identified 1,338 topological outliers—comprising approximately 1% of the analyzed population—defined as molecules that either act as generators for the most persistent topological features or reside in the most distal regions of the 12-dimensional space.

As detailed in Figure 6, these outliers exhibit distinct chemical signatures that separate them from the bulk population. They are characterized by significantly higher polarity, with a mean Topological Polar Surface Area (TPSA) of 49.94 $Å^2$ compared to 37.03 $Å^2$ for the bulk. Furthermore, outliers display a lower degree of saturation, evidenced by a fraction of sp^3 carbons (f_{sp^3}) of 0.485 versus 0.691 in the bulk, indicating a higher density of multiple bonds and aromatic systems. They are also significantly more hydrophilic (mean LogP of -0.34 compared to $+0.30$) and slightly lower in molecular mass (116.55 Da versus 122.75 Da).

Beyond their physical properties, these outliers are structurally atypical even within their own isomeric and scaffold families. The mean Tanimoto similarity between an outlier and its respective scaffold bulk is remarkably low, averaging just 0.179, with minimum similarities reaching as low as 0.00004. This indicates that the boundaries of the high-dimensional property manifold are anchored by molecules possessing highly unique local chemical environments and dense, unsaturated functional groups. These motifs create extreme electronic states that decouple from the global geometric axes, stretching conventional structure-property relationships to their limits.

4 Conclusions

In this study, we introduced a topological framework to rigorously quantify how macroscopic structural transitions rewire the fundamental physical relationships of small organic molecules. By isolating chemical architecture from stoichiometric determinism through the analysis of atom-count residuals, we mapped the complete 12-dimensional physical state space of the QM9 dataset. Utilizing persistent homology, we bypassed the limitations of traditional scalar correlation matrices to provide a global, coordinate-free measurement of manifold deformation across the Murcko scaffold hierarchy.

Our results demonstrate that macroscopic spatial confinement, specifically cyclization, induces a discrete and mathematically distinct reorganization of the joint property manifold. By computing the W_1 Wasserstein distances between the persistence diagrams of acyclic and cyclic populations, we established that this topological deformation significantly exceeds both intra-scaffold stochastic noise and stoichiometric baselines. This confirms that the observed manifold rewiring is a genuine structural signal driven by scaffold connectivity rather than compositional variance or simple linear shifts.

The physical mechanism underlying this topological reorganization is the fundamental decoupling of geometric and vibrational states. In acyclic chains, the lack of spatial confinement enforces tight couplings between global geometric extent and thermal response, as evidenced by the strong correlation between rotational constants and heat capacity. Upon cyclization, the topological constraints of ring formation restrict conformational freedom, relaxing these couplings and reshaping the joint property space.

Furthermore, our analysis of topological outliers revealed that the distal boundaries of this high-dimensional manifold are anchored by structurally atypical molecules. These outliers exhibit distinct chemical signatures—such as elevated polarity, increased hydrophilicity, and a lower fraction of sp3 carbons (f_{sp3})—indicating that the extremes of the chemical property space are defined by dense, unsaturated functional groups. These motifs generate extreme electronic environments that decouple from standard geometric axes, stretching conventional structure-property relationships to their limits.

Ultimately, this work highlights the utility of topological data analysis in chemical physics. By treating molecular property spaces as deformable high-dimensional manifolds, we can capture the exact structural consequences of spatial confinement. This framework provides a robust mathematical foundation for mapping complex structure-property relationships, demonstrating exactly how macroscopic architectural shifts fundamentally reshape the physical state space of organic molecules.